ChatGPT

Open-source AI models released in October 2025 (research summary)

October 2025 was remarkable for the release of high-performance open-weight models that researchers and developers can download and study. These models include reasoning-focused language models, vision-language models, safety guardrails and even a biology foundation model. They all have permissive licenses (MIT, Apache 2.0 or NVIDIA's Open-Model license) and most provide publicly available weights, training recipes and datasets. Table 1 summarises the ten most consequential open-source models released in October 2025.

Major open-source models

Model & release date	Parameters & architecture	License (all allow research and commercial use)	Benchmarks/position	No
MiniMax-M2 (MiniMax AI, 27 Oct 2025) [1]	Mixture-of-experts LLM with **230 billion** total parameters but only **≈10 billion** active during inference [1]; designed for efficient tool-use and agentic reasoning.	MIT	On the Artificial Analysis Intelligence index it is the **highest-ranked open-weight model (61 points)** and approaches proprietary models like GPT-5 [1]. It scored **77.2 on τ^2-Bench**, **69.4 on SWE-Bench Verified**, **66.8 on ArtifactsBench**, **75.7 on GAIA**, **44.0 on BrowseComp** and **65.5 on FinSearchComp-global** [1].	Op rel ba va too an inf Mo

Model & release date	Parameters & architecture	License (all allow research and commercial use)	Benchmarks/position	No
IBM Granite 4.0 Nano (IBM Research, 28 Oct 2025) [2]	Four small models (350M and **1.5 billion** parameters) using hybrid state-space and transformer architectures [2]. The models are designed for edge devices and local deployment.	Apache 2.0	Achieves **78.5 on IFEval (instruction following)** and **54.8 on BFCL-v3 (function calling)**; safety scores on SALAD and AttaQ exceed **90 %** [2]. The average benchmark score is **68.3 %** [3], leading other sub-billion LLMs.	Th pr ne pe ru GF th eff mo rea
DeepSeek-OCR (DeepSeek AI, 20 Oct 2025) [4]	Vision-language model combining a **380 M-parameter vision encoder** and a **3 B-parameter mixture-of-experts language decoder** with **570 M** active parameters [4]. Compresses long documents via image encoding to achieve 10× context compression.	MIT [5]	Achieves **97.3 % accuracy on the FoX benchmark** and sets new state-of-the-art on **OmniDocBench** while using far fewer vision tokens [4].	Op rel Hu de eff tex co bu an arc
Qwen3-VL-8B-Instruct (Alibaba, 15 Oct 2025) [7]	Dense transformer VLM with **≈8.77 B** parameters and native **256 K** context window (expandable to 1 M) [8]. Pre-trained on 36 trillion tokens and 2.5 million aligned image–text pairs [8].	Apache 2.0 [7]	On multimodal benchmarks it scores **69-70 on MMMU**, **≈77 on MathVista**, **≈896 on OCRBench**, and **≈96 % on DocVQA** [9].	Pr OC lar sp an un Re tra da
Qwen3-VL-8B-Thinking (Alibaba, 21 Oct 2025) [11]	Same architecture as the Instruct model (≈8.77 B parameters) but tuned to output longer chain-of-thought reasoning [9].	Apache 2.0	Slightly higher benchmark scores: **70-72 on MMMU**, **79-80 on MathVista**, **900–910 on OCRBench** and improved DocVQA accuracy [9].	En rea im so wi mo

Model & release date	Parameters & architecture	License (all allow research and commercial use)	Benchmarks/position	No
Kimi K2 Instruct (Moonshot AI, 24 Oct 2025 update)	Mixture-of-experts model with **1 trillion total parameters** and **32 B activated parameters** [12] ; supports 128K-256K context lengths and built-in agentic coding abilities.	Modified MIT license requiring that commercial products with over 100 M monthly users display “Kimi K2” [13] .	On coding and reasoning benchmarks the Instruct variant achieves **65.8 % pass@1 on SWE-Bench Verified (single attempt)** and excels at LiveCodeBench, OJBench and AIME 2025 tasks [14] [15] .	Op bo ins str co pe em co rea
OpenReasoning-Nemotron-32B (NVIDIA & collaborators, 28 Oct 2025) [16]	Derivative of Qwen2.5-32B; large reasoning model with **32 B parameters** and optional **1.5 B, 7 B and 14 B** variants [17] .	Creative Commons CC-BY-4.0 [18] .	Provides strong reasoning: the 32B model scores **64.3 on the Artificial Analysis Intelligence index** and sets new state-of-the-art across math and code benchmarks for its size [19] . With GenSelect inference it surpasses OpenAI O3 on AIME and coding tasks [20] .	Re tra an scr co im de
OpenFold3 preview (OpenFold consortium, 28 Oct 2025) [21]	Biomolecular foundation model replicating DeepMind’s AlphaFold 3. Parameter count not disclosed; trained on **over 300 000 experimental structures and 13 million synthetic structures** [22] .	Apache 2.0 [21]	Matches state-of-the-art open biomolecular models and approaches AlphaFold 3 in predicting monomeric RNA structures [23] .	Op rel co for nu lig pre de str mo

Model & release date	Parameters & architecture	License (all allow research and commercial use)	Benchmarks/position	No
Nemotron Nano 2 VL (NVIDIA, 28 Oct 2025) [24]	Vision-language model with **12 B parameters** using a hybrid **Mamba-Transformer** architecture. Supports FP4/FP8/ BF16 quantization and 128 K context [25] .	NVIDIA Open Model License [25]	Leads visual reasoning benchmarks: top performer on OCRBench V2 and improved accuracy across document intelligence and video captioning tasks [24] . Efficient Video Sampling (EVS) allows up to **2.5× higher throughput** without sacrificing accuracy [26] .	Ta mu op tra av Hu the da hig sa
Llama 3.1 Nemotron Safety Guard 8B V3 (NVIDIA, 28 Oct 2025) [27]	Safety guardrail model built on **Llama-3.1-8B** with 8.03 B parameters. Fine-tuned with LoRA on 386 K samples across 23 safety categories in nine languages [28] [27] .	NVIDIA Open Model License [27]	Achieves **84.2 % accuracy** in detecting harmful content across languages and categories [29] .	En mo pro res we Hu be Ne an sys

Comparison chart

The charts below compare parameter sizes and approximate benchmark scores for these models. Parameter sizes vary from <2 billion for IBM's Granite 4.0 Nano models to more than 30 billion for Kimi K2 and OpenReasoning-Nemotron-32B. Models specialising in safety and OCR (Safety Guard 8B V3 and DeepSeek-OCR) achieve high accuracy with comparatively small model sizes.

Parameter sizes of October 2025 open-source models

Approximate benchmark performance

Note: benchmark scores come from diverse evaluation suites (reasoning, code generation, OCR, safety). Scores are normalised for comparison but are not directly comparable across tasks. OpenFold3 and Nemotron Nano 2 VL have limited quantitative benchmarks available, so their bars are shorter or absent in the second chart.

Key observations

- **Diverse application areas:** the releases cover language models for reasoning (MiniMax-M2, OpenReasoning-Nemotron-32B), multimodal models for vision and document understanding (DeepSeek-OCR, Qwen3-VL, Nemotron Nano 2 VL), safety guardrails, small models for edge devices (IBM Granite 4.0) and biological structure prediction (OpenFold3). This demonstrates the breadth of open-weight research.
- **Permissive licenses enable reuse:** most models use Apache 2.0 or MIT licences; NVIDIA's models use their own Open Model License. These licences allow commercial use and fine-tuning, fostering community experimentation.
- **Mixture-of-experts vs. dense architectures:** high-performance models such as MiniMax-M2 and Kimi K2 adopt mixture-of-experts architectures to reduce active parameter count and inference cost while maintaining high accuracy [1] [12]. Dense transformers remain popular for smaller models like Qwen3-VL and IBM Granite 4.0.
- **Focus on reasoning and safety:** benchmarks highlight strong reasoning performance (MiniMax-M2, OpenReasoning-Nemotron-32B), competitive code-generation (Kimi K2), and advanced multimodal understanding (Qwen3-VL, DeepSeek-OCR). The release of Safety Guard 8B V3 reflects growing awareness of safety for open models [27].

These open-weight releases provide researchers with high-quality models, datasets and training recipes, accelerating innovation across fields from agentic AI to computational biology.

---

[1] MiniMax-M2 is the new king of open source LLMs (especially for agentic tool calling) | VentureBeat
https://venturebeat.com/ai/minimax-m2-is-the-new-king-of-open-source-llms-especially-for-agentic-tool

[2] [3] IBM's open source Granite 4.0 Nano AI models are small enough to run locally directly in your browser | VentureBeat
https://venturebeat.com/ai/ibms-open-source-granite-4-0-nano-ai-models-are-small-enough-to-run-locally

[4] DeepSeek drops open-source model that compresses text 10x through images, defying conventions | VentureBeat
https://venturebeat.com/ai/deepseek-drops-open-source-model-that-compresses-text-10x-through-images

[5] GitHub - deepseek-ai/DeepSeek-OCR: Contexts Optical Compression
https://github.com/deepseek-ai/DeepSeek-OCR

[6] GitHub - deepseek-ai/DeepSeek-OCR: Contexts Optical Compression
https://github.com/DeepSeek-AI/DeepSeek-OCR

[7] [10] Qwen/Qwen3-VL-8B-Instruct · Hugging Face
https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

[8] [9] Qwen3-VL-8B Instruct vs Qwen3-VL-8B Thinking: 2025 Guide
https://codersera.com/blog/qwen3-vl-8b-instruct-vs-qwen3-vl-8b-thinking-2025-guide

[11] GitHub - QwenLM/Qwen3-VL: Qwen3-VL is the multimodal large language model series developed by Qwen team, Alibaba Cloud.
https://github.com/QwenLM/Qwen3-VL

[12] [14] [15] GitHub - MoonshotAI/Kimi-K2: Kimi K2 is the large language model series developed by Moonshot AI team
https://github.com/MoonshotAI/Kimi-K2

[13] raw.githubusercontent.com
https://raw.githubusercontent.com/MoonshotAI/Kimi-K2/main/LICENSE

[16] [17] [18] [19] [20] nvidia/OpenReasoning-Nemotron-32B · Hugging Face
https://huggingface.co/nvidia/OpenReasoning-Nemotron-32B

[21] [23] GitHub - aqlaboratory/openfold-3: A fully open source biomolecular structure prediction model based on AlphaFold3
https://github.com/aqlaboratory/openfold-3

[22] OpenFold Consortium Releases Preview of OpenFold3: An Open-Source Foundation Model for Structure Prediction of Proteins, Nucleic Acids, and Drugs
https://www.businesswire.com/news/home/20251028507233/en/OpenFold-Consortium-Releases-Preview-of-OpenFold3-An-Open-Source-Foundation-Model-for-Structure-Prediction-of-Proteins-Nucleic-Acids-and-Drugs

[24] [26] [28] [29] Develop Specialized AI Agents with New NVIDIA Nemotron Vision, RAG, and Guardrail Models | NVIDIA Technical Blog
https://developer.nvidia.com/blog/develop-specialized-ai-agents-with-new-nvidia-nemotron-vision-rag-and-guardrail-models/

[25] nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-FP8 · Hugging Face
https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-FP8

[27] nvidia/Llama-3.1-Nemotron-Safety-Guard-8B-v3 · Hugging Face
https://huggingface.co/nvidia/Llama-3.1-Nemotron-Safety-Guard-8B-v3